Editorial: Computerlinguistik und Bibliotheken

Michael Franke-Maier[1] und Andreas Ledl[2]

[1] *Universitätsbibliothek der Freien Universität Berlin,*
franke@ub.fu-berlin.de
[2] *Universitätsbibliothek Basel,*
andreas.ledl@unibas.ch

Vor 50 Jahren, im Februar 1966, wies Floyd M. Cammack auf den Zusammenhang von „Linguistics and Libraries“ hin. Er ging dabei von dem Eintrag für „Linguistics“ in den Library of Congress Subject Headings (LCSH) von 1957 aus, der als Verweis „See Language and Languages; Philology; Philology, Comparative“ enthielt. Acht Jahre später kamen unter dem Schlagwort „Language and Languages“ Ergänzungen wie „language data processing“, „automatic indexing“, „machine translation“ und „psycholinguistics“ hinzu. Für Cammack zeigt sich hier ein Netz komplexer Wechselbeziehungen, die unter dem Begriff „Linguistics“ zusammengefasst werden sollten. Dieses System habe wichtigen Einfluss auf alle, die mit dem Sammeln, Organisieren, Speichern und Wiederauffinden von Informationen befasst seien. (Cammack 1966:73)

Bibliothekarinnen und Bibliothekare gehören hier selbstredend dazu, und dementsprechend werden als linguistische Segnungen für diesen Berufsstand die Verbesserung der Sprachkenntnisse sowie Unterstützung beim Katalogisieren und bei der Sacherschliessung aufgezählt. Im Zusammenhang mit Autoritätsdateien, Cross-Referenzen, Synonymen, Antonymen etc. bringt Cammack den Computer ins Spiel:

> „A cataloger who has spent much time at the task of maintaining a subject authority file or adjusting entries and cross references as new editions of classification schedules are issued can quickly see the implications of the linguists’ current project. A computer program which can handle the effectively infinite associational relationships in any given language would provide an ideal model for the equally complex problems of subject cataloging and indexing.“ (Cammack 1966:74)

Dabei seien weder der Einsatz des Computers noch sich verändernde Katalogisierungstechniken das eigentlich Entscheidende. Der originäre Beitrag der Linguistik für Bibliotheken bestehe vielmehr darin, dass sie nicht innerhalb eines Musters Punkte hinzufüge, sondern dieses Muster im Gegenteil erst entdecke und ermögliche, auf dieser Grundlage ein Klassifikationssystem zu entwickeln.

Cammacks Vision bestand darin, dass es eines Tages möglich sein werde, mittels computerlinguistischer Analysen und Persönlichkeitstests für jede(n) Studentin und Studenten bzw. jede(n) Wissenschaftlerin und Wissenschaftler eine individualisierte Ansicht des Bibliothekskatalogs bereitzustellen, die dank eines zu Grunde liegenden semantischen Netzwerks alle für

eine Person relevanten Materialien erfasst (Cammack 1966:74). Ausserhalb der Bibliothekswelt, d.h. für den kommerziellen Bereich, dürfte diese Vorstellung – man denke an Googles Suchmaschinentechnologie – bereits Realität geworden sein: Dieselbe Anfrage erzielt personenabhängig unterschiedliche Resultate. In den Discovery Engines und bibliografischen Fachdatenbanken der Bibliotheken sorgen die Filtermöglichkeiten über den persönlichen Account oder die Facetten zumindest für einen gewissen Grad an Individualisierung.

Folgt man der Entwicklungstendenz von Bibliotheken als Medienverwaltern über Datenvermittlern hin zu Informationsbeschaffern, so lassen sich grob fünf Anwendungsgebiete für automatische Verfahren differenzieren: Dokumentenretrieval, Passagen-Retrieval und Antwortextraktion, textbasierte Fragenbeantwortung und Informationsextraktion, automatisches Zusammenfassen sowie maschinelles Übersetzen (Hess und Clematide 2006). Auf die unterschiedlichen Anteile dabei und die begrifflichen Nuancen von Text and Data Mining (TDM), Natural Language Processing (NLP) und Computerlinguistik wollen wir an dieser Stelle nicht eingehen, das würde den Rahmen eines Editorials sprengen. Ein besonders häufig thematisierter Spezialbereich computerlinguistischer Methoden in Bibliotheken jedoch ist der Einsatz für die automatische Inhaltserschliessung – in den letzten Jahren oft gefordert, bei genauerem Blick jedoch nicht flächendeckend in der Bibliothekswelt angekommen, sondern mehr durch einzelne, wenn auch wichtige Projekte vertreten (Kasprzik 2014).[1]

Ob das, wie Alice Keller auf Basis einer Umfrage vermutet, an der „beträchtliche[n] Skepsis“ (Keller 2015:811) gegenüber automatischen Verfahren unter den befragten Expertinnen und Experten liegt – wie sie in ihrer Studie „Einstellung zur (automatischen) Sacherschließung in deutsch- und englischsprachigen Ländern“ attestiert –, ist zu bezweifeln: Die Erhebung und die Datenaufbereitung beanspruchen nirgendwo Repräsentativität. Insofern ist Kellers Umfrage als wichtiger Impuls zur richtigen Zeit zu begrüssen und sie könnte mit einer erweiterten und optimierten Fragestellung der möglichen Annahme, die automatische Inhaltserschliessung gegen die intellektuelle aufrechnen zu wollen, entgegnen. Nur ein Beispiel: Die Skepsis wird u.a. auf Basis der Auswertung der Frage „Gehen Sie davon aus, dass Computer den Menschen je ersetzen werden bei der Vergabe von Sachbegriffen oder Klassifikationen für Bücher?“ angenommen: „Auffallend ist der hohe Anteil an Befragten in beiden Sprachräumen (D 40,4%, E 27,9%), der davon ausgeht, dass Computer den Menschen nie ersetzen werden“. Kein Wort davon, dass im deutschsprachigen Raum 39,4% der befragten Personen (bei n=144) davon ausgehen, dass „Computer [. . .] den Menschen bis im Jahr 2025 ersetzen“ (Keller 2015:809) werden. Das ergibt sich, wenn man die unterschiedlichen Optionen, die Frage mit ja zu beantworten, summiert.

In Schürmanns Aufsatz „Sacherschliessung nach RDA“, in einer früheren Ausgabe dieser Zeitschrift erschienen (Schürmann 2015), werden maschinelle Verfahren im Kontext der internationalen Entwicklungen der letzten Jahre verortet. Wie kann ein künftiges Regelwerk die beiden Anforderungen von Bibliotheken, „Zugang zur Welt des Wissens und Schaufenster gepflegter Sammlungen“ (Schürmann 2015:78) zu sein, erfüllen: Für das erste ist sicherlich die Nutzung maschineller Verfahren – schon allein wegen der Menge an Publikationen – die ökonomischste Herangehensweise. Dahingegen ist bei dem Angebot einer unter einem bestimmten Fokus angelegten Sammlung an Publikationen die Kuratierung anhand hoher

[1] Der Artikel von Kasprzik konzentriert sich vor allem auf die automatische klassifikatorische Inhaltserschliessung des Zeitraums 2007-2012.

kongruenter inhaltserschliessender Metadaten von Vorteil. Wie ist letzteres mit technischen Mitteln zu erreichen?

Vor allem dann, wenn – nach RDA 23.4[2] – die Themenbeziehung als Kernelement definiert wird und damit „nicht eine jederzeit hinterfragbare Zusatzleistung, sondern integraler Teil des bibliothekarischen Kerngeschäfts der Erschließung“ (Stumpf 2015:1) ist? Stumpf beantwortet es eigentlich im Untertitel seines Vortrags zum Kerngeschäft Sacherschliessung: „Was gezielte intellektuelle Arbeit und maschinelle Verfahren gemeinsam bewirken können“. Und später schreibt er zur maschinellen Inhaltserschliessung:

> „Ihr Output ist aber an den Standards der intellektuell gesteuerten Erschließung zu messen, nicht umgekehrt. Sonst kann die Vision eines semantischen Netzes, zu dem Bibliotheken einen wesentlichen qualitativen Beitrag leisten wollen, nicht Wirklichkeit werden.“ (Stumpf 2015:13)

Jedem, der jetzt denkt, das klingt nach Skepsis, sei entgegnet: Hier liegt – im übertragenen Sinne – ein Heft vor Ihnen, in dem es um computerlinguistische Verfahren in Bibliotheken geht. Letztlich geht es um eine Versachlichung der Diskussion, um den Stellenwert der Inhaltserschliessung und die Rekalibrierung ihrer Wertschätzung in Zeiten von Mega-Indizes und Big Data. Der derzeitige Widerspruch zwischen dem Wunsch nach relevanter Treffermenge in Rechercheoberflächen vs. der Erfahrung des Relevanz-Rankings ist zu lösen. Explizit auch die Frage, wie oft wir von letzterem enttäuscht wurden und was zu tun ist, um das Verhältnis von *recall* und *precision* wieder in ein angebrachtes Gleichgewicht zu bringen. Unsere Nutzerinnen und Nutzer werden es uns danken.

Zu glauben, dass das günstig wird – und wir vielleicht mit einer kooperativen intellektuellen Inhaltserschliessung mit maschineller Unterstützung doch günstiger wegkämen, vor allem weil diese Mitarbeiterinnen und Mitarbeiter auch noch vieles andere erledigen könnten –, wäre arglos: Die Kosten der Entwicklung der künstlichen Intelligenz der Firma IBM, Watson – der bereits 2011 Jeopardy gewonnen hat, mittlerweile im Bereich Healthcare als clinical decision support system eingesetzt wird und bei Krebserkrankungen zu 99% identische Behandlungsmethoden vorschlägt wie Onkologen[3]; sicherlich auch für die Inhaltserschliessung einsetzbar wäre –, sind immens und können als pekuniärer Fixpunkt der Orientierung dienen: Die Fragen nach Buy-in und Minimum Bet gehen an die Managementebene.

Zurückkehrend zur Umfrage Kellers und zu Cammacks Vision des Computers als musterfindende Maschine kommen wir zum ersten Beitrag dieses Heftes, welcher sich eine Frage stellt, die Keller mit ihrer Fokussierung auf die menschliche Einstellungsebene umgeht und bei Cammack wahrscheinlich noch gar nicht im Fokus der Überlegungen war. Jene, die den Arbeitsbereich Lizenzierung betrifft und nicht erst seit dem Vortrag zu mehr Openness in Bibliotheken von Lohmeier (2014) Thema ist. Und zwar die Frage nach „Text and Data Mining“ im Zusammenhang mit Lizenzierung, vor kurzem in der Initiative Open Library Badge[4] u.a. als Kriterium für Openness aufgenommen.

[2] http://access.rdatoolkit.org/rdachp23-de_rda23-20028.html
[3] http://futurism.com/ibms-watson-ai-recommends-same-treatment-as-doctors-in-99-of-cancer-cases/
[4] https://badge.openbiblio.eu/

Kurz: Erlauben uns die derzeitigen Lizenzverträge überhaupt TDM[5] und falls nicht, welche Anstrengungen werden von den bibliothekarischen Vertreterinnen und Vertretern unternommen, um den anzustrebenden Zustand des „the right to read is the right to mine“ zu erreichen? CHRISTIAN WINTERHALTER gibt in seinem Beitrag „Licence to Mine?“ einen „Überblick über Rahmenbedingungen von Text and Data Mining und den aktuellen Stand der Diskussion“. Bemerkenswert ist sein Hinweis vor dem Hintergrund der bevorstehenden Reform des Urheberrechts, dass gerade „[z]um jetzigen Zeitpunkt auf den Lizenzweg zu setzen, [...] sowohl inhaltlich wie politisch als problematisch gelten“ kann. Ziel ist also gerade nicht den Lizenzweg zu beschreiten, sondern aktiv so Lobbyarbeit zu betreiben, dass TDM auf höchster Ebene in europäischer Gesetzgebung verankert wird.

Im darauffolgenden Artikel wird es zunächst historisch, bevor dann der Bogen in die Zukunft gespannt und aus der Sicht der Korpuslinguistik eine ganz neue Aufgabe für Bibliotheken skizziert wird: die Bibliothek als Korporathek! Zurückgreifend auf Vannevar Bush (1890-1974), der mit seiner Idee des „Memex“ als analoge Maschine zur Verarbeitung von Big Data die Grundlage zu den oben skizzierten digitalen Entwicklungen gelegt hat, geben NOAH BUBENHOFER und KLAUS ROTHENHÄUSLER in ihrem Aufsatz „‚Korporatheken‘: Die digitale und verdatete Bibliothek“ einen Einblick in die Arbeit der mit Computern betriebenen Korpuslinguistik, die daraus folgenden Herausforderungen für Bibliotheken und schliessen – passend zur Jahreszeit – mit einem Wunschzettel an Bibliotheken. Z.B. werden mehr Möglichkeiten des Open Data Processing und ein Angebot korpuslinguistisch nutzbarer Schnittstellen angeführt. Es wird auch hier wieder der Gedanke von mehr Openness aufgegriffen: Bibliotheken sollen neben den Metadaten auch die Volltexte als Material für die Forschung digital verfügbar machen.

Einen Aspekt, den Bubenhofer und Rothenhäusler dabei für die Korpuslinguistik gar nicht als so wichtig erachten – „ein perfekt reproduziertes digitales Abbild eines Originals“ –, beleuchten die beiden Autoren KONSTANTIN BAIERER und PHILIPP ZUMSTEIN im dritten Artikel „Verbesserung der OCR in digitalen Sammlungen von Bibliotheken“. Die Optical Character Recognition wird zunächst unter dem Aspekt von Erkennungsgenauigkeit betrachtet. Daraufhin werden typische Fehler gruppiert und es folgt eine Einordnung der computerlinguistischen Methoden bezüglich des gewinnbringenden Einsatzes im PostOCR-Verfahren. Bemerkenswert ist dabei der Hinweis auf den Widerspruch zwischen der Nutzung kommerzieller OCR-Produkte und nachhaltigen Lösungen. Dieser ist verbunden mit dem Fingerzeig, dass Bibliotheken eigentlich sehr viel mehr auf kooperative Eigenleistung setzen sollten, und damit wird indirekt die von Seeliger angestossene Berufsbild-Diskussion mit „stärkere[m] IT-Profil“ (Seeliger 2015:265) befürwortet.

Dass diese Diskussion – eigentlich dauerhaft – geführt werden muss und auch in vielen Bibliotheken bereits nicht nur ernst genommen, sondern erfolgreich umgesetzt wird, zeigt sich beispielhaft an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW). Der vierte Artikel stammt von zwei Mitarbeitern dieser wirtschaftswissenschaftlichen Forschungsbibliothek, namentlich MARTIN TOEPFER und ANDREAS OSKAR KEMPF. In ihrem Aufsatz „Automatische Indexierung auf Basis von Titeln und Autoren-Keywords – ein Werkstattbericht“ geht es um Methoden der Nachnormierung von Autoren-Stichwörtern, die

[5]Diese Fussnote dient als ergänzender Lektürehinweis: Ein weiterer aktueller Beitrag zu TDM von Bastian Drees ist mit einem anderen inhaltlichen Zuschnitt gerade in der Zeitschrift *Perspektive Bibliothek* erschienen.

in der Masse einen unüberschaubaren heterogenen Pool erzeugen. Wie kann einer solchen inkongruenten Erschliessungssituation entgegnet werden? Toepfer und Kempf liefern einen klassischen computerlinguistischen Ansatz, der bereits erschlossene Referenzindexate mit einem automatischen Verfahren, das an der ZBW entwickelt wurde, erneut erschliesst. Die anschliessende Berechnung von *recall, precision* und *f1-Wert* gibt Aufschluss darüber, wie gut das Verfahren arbeitet – Resümee: es scheint auch für grosse Datenmengen eine aussichtsreiche Methode zu sein.

PETER KRAKER, CHRISTOPHER KITTEL und ASURA ENKHBAYAR beschäftigen sich im vorletzten Beitrag des Heftes mit der Bereitstellung eines visuellen Web-Interface ebenfalls für grosse Datenpools, zeitnah zur Veröffentlichung eines Papers in PudMed und DOAJ[6]: Ziel ist, mit so genannten Wissenslandkarten Publikationen zu gleichen Themen zu clustern und ggf. bisher versteckte Beziehungen zwischen Dokumenten aufzuzeigen. In ihrem Aufsatz „Open Knowledge Maps: Creating a Visual Interface to the World's Scientific Knowledge Based on Natural Language Processing" beschreiben die Autoren ihr konkretes Projekt, die Technik dahinter sowie die künftigen Vorhaben zur Optimierung der Ergebnismengen. Vorbildlich ist dabei die Haltung, dass das Projekt „completely open" sein wird, sich in das offene Ökosystem durch Anbindung an andere bereits etablierte, offene Software integrieren wird und als kollaboratives Werkzeug[7] unterschiedliche Communities zusammenbringen will.

Das Heft schliesst mit einem Artikel von CLEMENS NEUDECKER und GEORG REHM zu „Digitale[n] Kuratierungstechnologien für Bibliotheken" und gibt damit Einblick in ein Projekt, welches im September 2015, vom Bundesministerium für Bildung und Forschung gefördert, startete. Als Grundidee wird die Unterstützung von „komplexe[n], von Redakteuren und Wissensarbeitern durchgeführte[n] digitale[n] Kuratierungsprozesse[n] durch Sprach- und Wissenstechnologien" verfolgt. Die Liste der dabei eingesetzten Technologien ist überwältigend und fast enzyklopädisch für die Computerlinguistik: Informationsextraktion, Named Entity Recognition, Temporale Analyse, Geolokalisierung, Annotation mit allgemeinen Metadaten, Clustering, Klassifikation, Sentiment-Analyse, Textgenerierung, Semantic Storytelling, maschinelle Übersetzung und mehrsprachige Linked Data. Was will man mehr?

Letztlich nur eine anregende Lektüre, die wir Ihnen beim Lesen des Heftes wünschen – in der Hoffnung, dass der ein oder andere Impuls auch in Ihrer Institution fruchtbar umgesetzt wird.

Ihr
Michael Franke-Maier (Gast-Herausgeber)
und
Andreas Ledl

[6] http://openknowledgemaps.org/news.php
[7] Zu den kollaborativen Funktionalitäten gibt es auch ein schönes Video unter https://vimeo.com/188647919.

Literatur

Cammack, F. M. (1966). Linguistics and Libraries. In: *Stechert-Hafner Book News* XX.6, S. 73–75.

Drees, B. (2016). Text und Data Mining: Herausforderungen und Möglichkeiten für Bibliotheken. In: *Perspektive Bibliothek* 5.1, S. 49–73. DOI: 10.11588/pb.2016.1.33691.

Hess, M. und Clematide, S. (2006). *Computerlinguistik in Information und Dokumentation. Kurs für wissenschaftliche Bibliothekare*. Universität Zürich: Institut für Computerlinguistik. URL: https://files.ifi.uzh.ch/cl/siclemat/talks/zb/zb.pdf.

Kasprzik, A. (2014). Automatisierte und semiautomatisierte Klassifizierung – eine Analyse aktueller Projekte. In: *Perspektive Bibliothek* 3.1, S. 85–110. DOI: 10.11588/pb.2014.1.14022.

Keller, A. (2015). Einstellung zur (automatischen) Sacherschließung in deutsch- und englischsprachigen Ländern. In: *Bibliotheksdienst* 49.8, S. 801–813. DOI: 10.1515/bd-2015-0095.

Lohmeier, F. (2014). *Leitbild Openness – Bibliotheken als Wächter für den (dauerhaft) freien Zugang zum Wissen*. Vortrag, 103. Deutscher Bibliothekartag in Bremen. URL: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0290-opus-16542.

Schürmann, H. (2015). Sacherschliessung nach RDA. In: *027.7 Zeitschrift für Bibliothekskultur* 3.2, S. 74–80. DOI: 10.12685/027.7-3-2-64.

Seeliger, F. (2015). Informatiker, Journalisten, Erzieher. Anforderungen an den bibliothekarischen Berufsstand von heute und morgen. In: *b.i.t.online* 18.3, S. 264–265. URL: http://www.b-i-t-online.de/heft/2015-03-nachrichtenbeitrag-seeliger.pdf.

Stumpf, G. (2015). *„Kerngeschäft“ Sacherschließung in neuer Sicht: Was gezielte intellektuelle Arbeit und maschinelle Verfahren gemeinsam bewirken können*. Vortrag, Fortbildungsveranstaltung für Fachreferentinnen und Fachreferenten der Politikwissenschaft und Soziologie, 22./23. Januar 2015 in Berlin. URL: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:384-opus4-30027.